DOCUMENT RESUME

ED 072 106 TM 002 347

AUTHOR Porter, Andrew C.
TITLE Some Design and Analysis Concerns for Quasi-Experiments such as Follow Through.
PUB DATE Aug 72
NOTE 25p.; Paper presented at the American Psychological Association meetings, 1972

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Analysis of Covariance; *Evaluation Methods; *Mathematical Models; *Program Evaluation; *Statistical Analysis
IDENTIFIERS *Project Follow Through

ABSTRACT

The basic design for the national evaluation of the Follow Through program is presented, and some of the related issues of analysis are considered. The design, as it now stands, presents many difficulties for analysis. These analysis issues are seen to include the following: (1) What should be the unit of analysis?; (2) How is the effect of a Follow Through approach when compared with its control estimated?; and (3) How is the relative effectiveness of the various approaches estimated? Several different strategies that have been suggested for use in quasi-experiments in attempts to control variables that are confronted with treatments are discussed. The potential confounding variables are classified into two categories: (1) systematic differences in the dependent variable dimensions that are present in the units of analysis at the outset of program participation; and (2) systematic differences that occur in the dependent variable dimensions during program participation which are not a function of program participation. The appropriate method for attempting to control confounding variables in the evaluation of Follow Through appears to be a combination of strategies employing both matching and estimated true score ANCOVE. (DB)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

ED 072106

SOME DESIGN AND ANALYSIS CONCERNS FOR QUASI-EXPERIMENTS SUCH AS FOLLOW THROUGH

Paper presented at the 1972 APA meetings

TM 002 147

Andrew C. Porter

Michigan State University

August 1972

FILMED FROM BEST AVAILABLE COPY

The purpose of the present paper is to present the basic design for the national evaluation of the Follow Through program, and to consider some of the related issues of analysis. In so far as the design for Follow Through is representative of other large scale quasi-experiments, the discussion of analysis issues may have some generality beyond the Follow Through evaluation.

Briefly stated, Follow Through is a community action program consisting of several different approaches designed to improve the life chances of children in families who are living at a poverty level. In order to facilitate early evaluation of the various approaches, some assumptions have been made about what currently available data are predictive of life chances. Although there is far from total agreement on what currently available data are predictors of life chances or on their relative importance, a partial list might be: cognitive, affective, social, and physical characteristics of the children that participate; their parent's attitudes toward them and school, and their parent's involvement in the educational process; teacher attitudes and behaviors. Although some of the analysis issues may be common across these broad categories of assumed predictors of life chances, discussion in the present paper will be centered on analyzing the cognitive and affective characteristics of the children. The main questions to be addressed in the analysis are:

1) What differences, if any, are there between a Follow Through approach and its non-Follow Through comparison?
2) What differences, if any, are there among the various Follow Through approaches?
3) Does a Follow Through approach have different effects across types of communities and children?

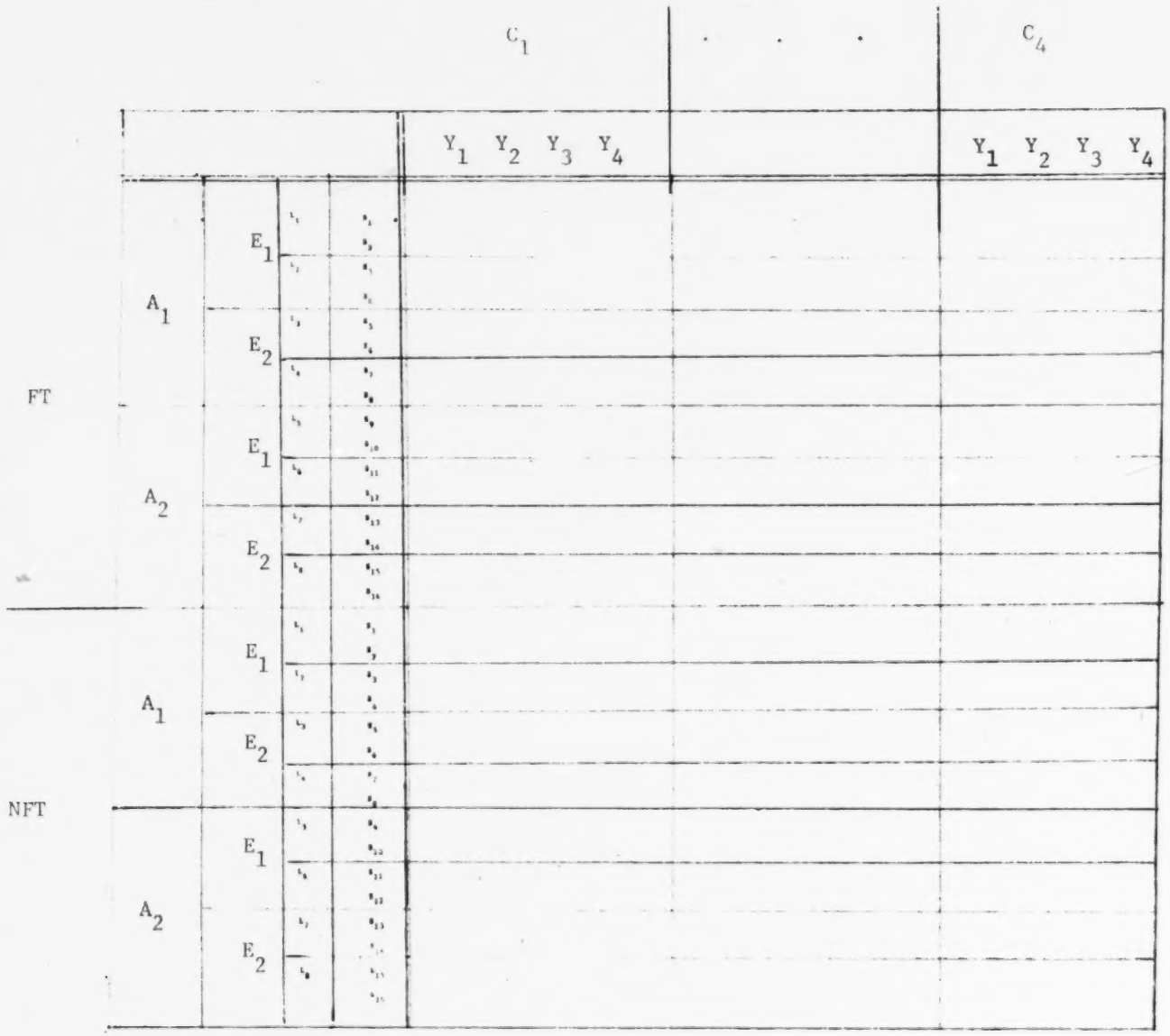
The general strategy for addressing the evaluation questions has been to have each Follow Through approach implemented in several different locations or projects. For every school where a Follow Through approach has been implemented a matched non-Follow Through school has been identified. The matched non-Follow Through schools were identified by the local Follow

Through project leaders who considered such dimensions as ethnic group composition and socioeconomic status. The Follow Through program is a four year experience for children who enter school at the kindergarten level and a three year experience for children who enter at the first grade level. Children are tested in the Fall and Spring of their first year and then each Spring for their remaining years in the program. A group of children entering the Follow Through program in a given year is called a cohort. The national evaluation will consider four cohorts of children with the first cohort starting in the Fall of 1969 and completing the program at the end of their third grade year. The last cohort will start this Fall, 1972.

One representation of the basic design for the national evaluation of Follow Through is presented in Figure 1. On the verticle dimension of the data matrix is Follow Through (FT) versus non-Follow Through (NFT). Two Follow Through approaches (A) are represented for illustration purposes. Presently there are twenty different sponsors and in some respects each sponsor represents a different approach. The dimension, E, has been included in the data matrix to represent one of the many dimensions that might interact with the FT - NFT and/or A dimensions. For purposes of illustration let E denote ethnic composition of a school. Two locations (L) are indicated for each sponsor and within each location two schools (S) are represented. Across the top of the data matrix is the dimension of four cohorts (C) and for each cohort four years (Y) or grade levels (kindergarten through third grade). Actually the design includes yet another dimension for the entry level of children, i.e. kindergarten or first grade. Figure 1, represents the design for children who enter the program in kindergarten.

Although Figure 1 is not the actual design it will serve as a base for discussion. In Figure 1, locations are nested within Follow Through approaches, i.e. the locations are different for each approach. Ideally, these locations would be randomly selected from an appropriate population and then randomly assigned to approaches. Such random assignment helps to avoid systematic differences in locations across approaches that might

Figure 1



FT, NFT	denotes Follow Through and its comparison
A	denotes Follow Through approach or sponsor
E	represents one of many variables that are of interest because they may interact with Follow Through approach
L	denotes location of a project
S	denotes matched pairs of schools
C	denotes cohort
Y	denotes year in school, i.e. kindergarten - 3rd grade

become competing hypotheses to any observed differences on the dependent variables among approaches. Further, schools are shown to be nested within both approaches and locations but crossed with FT - NFT. The crossing is not because the same school contains both FT and NFT children for the evaluation (although that has occasionally happened), but rather to illustrate that schools are in matched pairs, i.e. S_1 under FT and S_1 under NFT represent two schools in a matched pair. The matching of schools is also why approaches and locations are crossed with FT - NFT in Figure 1. Ideally then, matched pairs would be formed initially and then one school randomly assigned to FT and one to NFT to avoid systematic differences at the outset between FT and NFT schools. Given the above prescribed randomization and the balanced and fully crossed nature of the data matrix in Figure 1, the analysis and subsequent testing of hypotheses would be rather straight forward.

The actual design of the Follow Through evaluation represents several deviations from that presented in Figure 1 and does not include appropriate random selection and assignment procedures. Locations were judgementally chosen for participation in Follow Through, and following their selection the project leaders in the locations decided upon which approach was to be implemented. In each location a school (or schools) was chosen for implementation of the Follow Through program and later an attempt was made to locate a similar school for comparison. The result is that prior to their participation in the program Follow Through children differed in systematic ways from non-Follow Through children. However, because of the attempt to select matched schools for purposes of comparison, these differences are somewhat under control On the other hand, no attempt was made to match the schools for one approach with the schools for another approach. This has created rather serious confounding of type of students and perhaps even type of NFT program across approaches. For example, in Cohort I, one sponsor has all white children while another sponsor has 83% Black. (U.S. Office of Education, 1972, p.12) This problem continues across cohorts. In terms of Figure 1, this means that E and A are not completely crossed. A further deviation from the design depicted

in Figure 1 is that some approaches and locations are crossed rather than completely nested. This partial crossing further complicates attempts for a single overall analysis.

The longitudinal nature of the design presents another series of complications. Not all schools are completely crossed with cohorts and/or grades. For example, of the 70 entering kindergarten sites for Cohort I only 20 had testing at the end of first grade and 29 at the end of second grade (U.S. Office of Education, 1972, p. 10). However, the intention is that many of the original 70 that didn't receive testing in the intermediate years will be tested in the Spring of 1973 which is their final year in the program. A further complication for any longitudinal analysis of these data is that the test battery has not remained the same. First, the tests have changed across cohorts in an attempt to improve quality of the test battery, particularly in the non-cognitive domain. Second, even if the test battery had remained the same across cohorts, there would have been some changes across years for a cohort. Because of such factors as changes in item difficulty tests appropriate for one age level are frequently not appropriate for another age level.

In summary then the design deviates in many respects from the ideal. These deviations may stem in large part from Follow Through's early purpose of being a comprehensive service program rather than an experiment to test various approaches. Whatever the reasons, the design, as it currently stands, presents many difficulties for analysis. Many of these difficulties would have been alleviated by proper random selection and assignment procedures. Although I feel such procedures can be accomplished in the evaluation of programs such as Follow Through and that they are worth the effort, these arguments have already been presented and will not be repeated here (Porter, 1969; Campbell and Erlebacher, 1970). On the other hand, some of the difficulties with the longitudinal aspects of the design were unavoidable, such as tests not being crossed with years because a test for one age is inappropriate for another age. The lack of intermediate testing was a result of financial restrictions and some of the changes in the test battery across cohorts were a result of insuffic-

ient time allowed by the U.S. Office of Education for the development of a test battery prior to the start of the first experimental cohort. The question became, should the original test battery be continued to facilitate analyses across cohorts or should the test battery be improved?

Because of the lack of random assignment of schools to FT - NFT and of locations to approaches and because of the incomplete nature of the longitudinal data, there is no single overall analysis that can use all of the data in testing the hypotheses of interest. Nor is there likely to be a single best strategy for multiple analyses. Rather there seems to be a need for using several alternative strategies each having some weakness and some unique strengths. If the same conclusions are reached across multiple analysis strategies, each based on different assumptions, then we can have greater confidence in the conclusions than if only a single analysis strategy had been employed. This idea is similar to that of the replication of findings across cohorts strengthening our conclusions.

The analysis issues involved in the national evaluation of Follow Through seem to fall into five broad categories:

1) Given the lack of random assignment, how can the actual effects of the various approaches be estimated without contamination from variables confounded with FT - NFT and/or the approaches?

2) How can the data be used to investigate interactions of various dimensions with approaches, given that schools were not selected with these dimensions in mind?

3) What analyses will make best use of the longitudinal nature of the data?

4) How should the analyses deal with the multiple outcome measures?

5) How can the data be analysed with maximum statistical power?

Attempting to deal with all five categories in detail if not too ambitious for a single paper, is at least too ambitious for this single author. Instead the following discussion will focus on some of the analysis issues that have been of particular interest to me during my

involvement with the evaluation of Follow Through and to some extent will cut across the five categories.

One of the most basic analysis issues, and one that really precedes consideration of the five previously mentioned categories, is what should be the unit of analysis? Should the unit of analysis be an individual child, a classroom of children, a school, or a location? One answer is that when inferential statistics are to be employed, the unit of analysis should be the same as the experimental unit, i.e. the smallest group of children that receives a FT or NFT experience independent of all other groups. This requirement is based on the assumption of independence which underlies all inferential procedures either parametric or nonparametric. However, this answer is complicated in educational studies because there are usually degrees of independence. For example, students in a single classroom are more independent of each other when classroom discussion is discouraged than when classroom discussion is encouraged. In the Follow Through evaluation, classrooms in a school are probably more independent of each other than students in a classroom. Still, schools in a location are probably more independent of each other than are classrooms in a school, and locations or projects are probably more independent than schools in a project.

It could perhaps be argued that classrooms in a grade school are sufficiently autonomous to be considered the units of analysis, but this choice would appear to complicate the longitudinal analyses of the data. Classrooms are generally groups of students that exist for a single grade so that they are not crossed with years for a cohort. However, nesting classrooms within years of a cohort would not take into account that the students comprising the classrooms are crossed with years. Treating classrooms as nested within years is the type of mistake that will, in general, make an analysis too conservative in the sense of statistical power. Further, classrooms are obviously not crossed with cohorts. The choice of school as the unit of analysis gets the unit of analysis one step closer to independence and solves the problem of how to represent in the analysis the design fact that students are crossed with years for a cohort. As shown in Figure 1, schools are crossed with both cohorts and years.

One might ask why not use location as the unit of analysis rather than school since this would be yet one step closer to independence? Perhaps location should be the unit of analysis. If Follow Through is considered an experimental study of different approaches on only a sample of locations, with the intent that the better approaches will be implemented in new locations, then the choice between school and location is not important. This is because an analysis which wishes to generalize beyond its sample of locations will treat locations as a random factor employing the Cornfield - Tukey bridge argument (Cornfield and Tukey, 1956). If locations are a random factor in the design in Figure 1, the correct error variance for testing hypotheses about FT - NFT and approaches will be variability among locations regardless of whether schools or locations were originally designated as the units of analysis. On the other hand, if locations are considered a fixed factor, either schools must be the unit of analysis or at least one interaction involving locations must be assumed equal to zero in order that the design afford tests of the hypotheses about FT - NFT and approaches. These statements are based on inspection of the expected mean squares for the sources of variation in the design under each condition (Wright, 1969).

Some people have argued that although the unit of analysis and the experimental unit should be synonomous to facilitate inference, such a choice will prevent the investigation of treatment by child characteristic interactions which are of interest. From an evaluation point of view, however, such interactions may not be of interest. Almost all Follow Through sponsors are using classroom oriented approaches. If one approach works better with black children and another approach works better with white children, then what are the implications for integrated classrooms? Should both approaches be used in an integrated classroom? Such a decision would not be based on data from the evaluation since the interaction was observed for situations where children were in classrooms receiving only one approach. It seems more appropriate to investigate treatment interactions with variables defined on classroom composition, such as percent of white children in the class. Where school is the unit of analysis the

variables should be defined on the composition of only those classrooms in the school which are participating in the study. A treatment by classroom composition interaction suggests that approaches should be selected for classrooms at least in part on the basis of classroom composition.

If schools are to be the units of analysis then the testing program for Follow Through should focus on testing in as many schools per approach as is feasible. This is so that the design has sufficient degrees of freedom to support multivariate analyses and to facilitate statistical power. Implementing Follow Through programs in more schools and then testing in those schools would represent a great expense. However, there are currently schools with Follow Through programs that are not being tested, particularly in the intermediate years of a cohort. If schools were the units of analysis, an observation on a school could be based on a sample of the children or a sample of the classrooms from that school. If the samples were taken randomly, no bias would be introduced into the data. Sampling children and/or classrooms from schools might not represent much of a savings in testing dollars for group administered tests, but it might represent a considerable savings for individually administered tests. Perhaps the dollars saved by not testing all children in a school could be used to do testing in additional schools.

An argument advanced for testing all children in a school is based on the problem of heavy attrition in Follow Through programs. If not all children are originally tested, then it is feared there may be none of the originally tested children left at the end of third grade. If children were the units of analysis this would seem to represent a more severe problem than with schools as the units. Certainly the schools will still be around at the end of third grade for a cohort. The random sample of children originally tested will still be an unbiased estimate of what children in the program were like at the outset, and a random sample of children in the program at the end of third grade will be an unbiased estimate of what children in the program were like at the conclusion. The analysis should include attrition rate per school as an additional outcome measure. Using the strategy of limiting analyses to those

students who stay in the program from beginning to end would seriously jeopardize the generality of the findings, particularly since the interest is in populations which have heavy attrition. Such a strategy should tend to make Follow Through look stronger than it really is with the entire population of interest.

The most important and most difficult analysis issues in the national evaluation of Follow Through are:

1) how to estimate the effect of a Follow Through approach when compared to its control,

and 2) how to estimate the relative effectiveness of the various approaches.

As has already been stated the design attempted to match Follow Through schools to non-Follow Through schools, ie. in Figure 1 pairs of schools and FT - NFT are crossed. However, there was no attempt to match schools in one approach with schools in another. This is illustrated in Figure 1 by having both schools and locations nested within approaches. For a given approach an FT - NFT comparison can be made on students from similar geographic locations and roughly similar ethnic composition. This is not to say that the effort to match schools was totally successful and that unbiased estimates of Follow Through effects for each approach are straight forward, but rather that the comparison groups and their school programs were at least roughly comparable prior to Follow Through.

Rough comparability is not at all the case when the interest is in comparing one approach to another. One approach might be implemented primarily in the south and another primarily in the north-east or one might be implemented in schools comprised of primarily black children and another in schools comprised of primarily white children (U.S. Office of Education, 1972). Such gross initial differences from one approach to another make direct comparisons of approaches questionable. One suggestion has been to compare the FT - NFT difference for one approach to the FT - NFT difference for another approach thus making an indirect comparison of approaches. Unfortunately, this method of indirect comparison leaves

approach by school characteristic interactions confounded with approaches. Imagine that the FT - NFT comparison for an approach implemented with schools comprised primarily of black children shows a greater difference in favor of Follow Through than the FT - NFT comparison for an approach implemented with schools comprised primarily of white children. The temptation of one employing the indirect comparison of approaches strategy might be to conclude that the approach with the greater FT - NFT difference is the better approach. However, this conclusion might be totally unwarranted because of an approach by ethnic composition interaction. If both approaches were implemented in schools comprised primarily of black children their FT - NFT differences might be equal. Smith and Bissel (1970) in their reanalysis of the data from the Westinghouse-Ohio University evaluation of Head Start suggest that there may have been such an interaction in the Head Start data. Another serious threat to the validity of conclusions reached using the indirect comparisons strategy is that approach and quality of NFT comparison school programs might be confounded. An approach whose NFT comparison schools typically receive large amounts of Title I funds may be at a disadvantage when compared to an approach whose NFT comparison schools receive relatively little Title I funds.

Not only does the design strategy of matching FT schools to NFT schoools not facilitate between approach comparisons, it has made the testing program more expensive than it otherwise might have been. Had it been possible to identify a population of locations and schools for possible Follow Through programs and then randomly assign locations to approaches, comparisons of approaches could have been done directly. Further, such a strategy would have required only a single group of NFT locations for a control group rather than a separate control for each approach. In terms of Figure 1, rather than having to test in 16 NFT schools in all 8 Follow Through locations, testing could have been done in only four schools in two randomly equivalent locations.

Several different strategies have been suggested for use in quasi-experiments in an attempt to control variables which are confounded with

treatments and thus offer rival hypotheses for any differences (or lack of differences) that are found in the data. The remainder of the present paper considers several such strategies as they relate to the Follow Through design. In considering these strategies it will be helpful to classify the potential confounding variables into two categories:

1) systematic differences in the dependent variable dimensions that are present in the units of analysis at the outset of program participation,

2) systematic differences that occur in the dependent variable dimensions during program participation which are not a function of program participation.

Category one differences are probably best reflected in pretest differences. The second category of differences are less straight forward to estimate. What if the average home environment of FT children is inferior to that of the NFT children during their period of participation or lack of participation in Follow Through? Home environment would seem to represent a treatment which may well affect the dependent variable dimensions of the evaluation. In fact, the home environment treatment might be more potent than the Follow Through treatment, thus making Follow Through appear detrimental when in fact the opposite might have been the case.

A strategy that controls many of the potential confounding variables in quasi-experiments has been labeled by Campbell and Stanley (1963) as the multiple time-series design. Implementing the multiple time-series design for the Follow Through evaluation would require augmenting Figure 1 with several measures on children prior to their going to school with the measures being equally spaced over time. For each cohort then there might be eight years, four spring testings prior to entry into kindergarten and the four spring testings represented in Figure 1. Such a design affords an estimate of the trends over time for each dependent variable prior to program participation. If the trend for FT children changed in a way different from the trend for NFT children it would suggest a Follow Through effect. It should be noted, however, that even this design does not control for possible school characteristics by

approach interactions that might be confounded when comparing approaches.

Obviously this strategy cannot be implemented for the dependent variables defined by the current testing program because none of the first four data points are available. I'm not even sure that such a strategy could be implemented for use on a cohort to start kindergarten four or five years from this fall. The problem is that tests appropriate for students in kindergarten may well not be appropriate for preschoolers nor for children in fourth grade. Any change in metric across the eight years would interrupt the estimation of trends.

One of the most common strategies for controlling variables confounded with treatments is analysis of covariance (ANCOVA). ANCOVA has primarily been identified with controlling for the first category of confounding variables, i.e. those present at the outset of program participation; however, it may also control for some special cases of confounding which fit the second category. Before continuing the discussion, a very brief review of the ANCOVA strategy will be helpful. Rather than testing for equality of populations means directly on the outcome variable, say Y, as analysis of variance of the posttest would do, ANCOVA tests for the equality of population means after having adjusted for differences on a covariable, say X. For purposes of illustration, consider a one-way model. The ANCOVA adjusted means are

$$\mu'_{Y.j} = \mu_{Y.j} - \beta_{Y.X} (\mu_{X.j} - \mu_{X..}) \quad ,$$

where the prime indicates adjusted, j denotes treatment group, and $\beta_{Y.X}$ denotes the slope of the within treatment group regression line for predicting Y from knowledge of X. Actually in classical ANCOVA, covariables are fixed and so the means on X are written as sample means. I have represented them as population means instead, because the Follow Through design does not provide any fixed covariable, and because De Gracie (1968) has shown that classical ANCOVA procedures when applied to data with a random covariable provide valid test statistics. The difference between two adjusted means is then

$$\mu'_{Y_{1.}} - \mu'_{Y_{2.}} = \mu_{Y_{1.}} - \mu_{Y_{2.}} - \beta_{Y.X}\,(\mu_{X_{1.}} - \mu_{X_{2.}}) \ ,$$

or in words it is the difference between the posttest means after subtracting $\beta_{Y.X}$ times the difference between the pretest means. Unfortunately $\beta_{Y.X}$ is not the correct multiplier for removing covariable differences since it represents the slope of the regression line defined on the observed variables rather than their latent true parts free from errors of measurement (Lord, 1960; Porter, 1967). It is not difficult to show that the desired slope is equal to the reciprocal of the reliability of the covariable times the least squares slope of the observed variables, i.e.,

$$\beta_{Y_T.X_{T.}} = 1/\rho_{XX'}\,\beta_{Y.X} \ ,$$

where the subscript T indicates a variable free from errors of measurement and $\rho_{XX'}$ denotes the reliability of the observed covariable X. For this reason classical ANCOVA is not an appropriate strategy unless the covariables are prefectly reliable.

I have developed a modification of classical ANCOVA which deals with the problem of having covariables that are fallible, i.e. contain errors of measurement (Porter, 1967, 1971; Campbell and Erlebacher, 1970). My modification uses exactly the same computational procedures as classical ANCOVA after having first substituted an estimated true score covariable for the observed covariable. The estimated true score covariable for a one-way model is defined

$$\hat{T}_{ji} = \bar{X}_{j.} + \rho_{XX'}\,(X_{ji} - \bar{X}_{j.}) \ ,$$

where $\hat{T}_{ji}$ denotes the estimated true score for the ith unit of analysis in the analysis in the jth treatment group. For more complex designs the estimated true score covariable would follow the same form except that the observations would be deviated from the respective cell means. The important properties of an estimated true score covariable are that

it is a linear transformation of the fallibly measured covariable and:

1) has the same treatment group and grand means as the fallibly measured covariable and the unobserved true covariable,

2) has the same correlation with the dependent variable as does the fallibly measured covariable,

and 3) the slope, $\beta_{Y\hat{T}}$ is equal to the desired slope of the latent true variables $\beta_{Y_T \cdot X_T}$ (Porter, 1971).

To the extent that the covariables used reflect initial group differences that are predictive of posttest differences, my estimated true scores ANCOVA is a useful approach for controlling confounding variables falling in category one. As mentioned earlier pretests are probably the best covariables for controlling initial differences, however, one of the strengths of the procedure is that the covariables need not be pretests.

In special situations estimated true scores ANCOVA may also control for confounding variables of the type falling in category two. The special situation where this is true has been labeled by Campbell (1971) as the fan spread situation. Briefly stated, the fan spread hypothesis is that a dependent variable dimension at the time of posttest is a linear transformation of the dependent variable dimension at the time of pretest except for any treatment effects. Consider a treatment group and its control that are two points different on a pretest and 4 points different on the posttest and further that the treatment actually had no effect. We would label the two point change in difference as the result of some confounding variable falling in category two. For example, a differential maturation rate that led to the initial two point difference between groups might have continued during the study and led to an eventual four point difference. The fan spread hypothesis says that the 2 point increase in difference will be accompanied by a four fold increase in variance, i.e., the posttest scores are two times the pretest scores.

Recalling now the difference between adjusted means for classical ANCOVA and substituting the slope defined on the latent true variables for the slope defined on the observed variables as my estimated true

score ANCOVA does, we have

$$\mu'_{Y_{1.}} - \mu'_{Y_{2.}} = \mu_{Y_{1.}} - \mu_{Y_{2.}} - \left(1/\rho_{XX}\right) \beta_{Y.X} \left(\mu_{X_{1.}} - \mu_{X_{2.}}\right) .$$

Since $\beta_{Y.X} = \rho_{XY}\, \sigma_Y/\sigma_X$, where σ denotes standard deviation, the $1/\rho_{XX}$ factor corrects the correlation ρ_{XY} for attenuation due to both X and Y and makes it equal to one. Therefore, the product of $1 / \rho_{XX}$ times $\beta_{Y.X}$ becomes the ratio σ_Y/σ_X. For our example the difference in adjusted means is then

$$\mu'_{Y_{1.}} - \mu'_{Y_{2.}} = 4 - \sigma_Y/\sigma_X \quad (2) .$$

But σ_Y/σ_X is equal to 2 given the fan spread hypothesis and so the difference in adjusted means is zero. The finding of no difference between treatments is consistent with what we know to be correct for our hypothetical example and illustrates the ability of estimated true scores ANCOVA to control for confounding which follows the fan spread hypothesis. Unfortunately there may be confounding in the Follow Through data that is not reflected in initial pretest differences and does not conform to the fan spread hypothesis. To the extent that this is so, estimated true scores ANCOVA is not a totally satisfactory strategy.

Another strategy for controlling confounding variables in quasi-experiments has recently been suggested by Campbell (1971). The strategy involves calculating the correlation between treatment group membership and the dependent variable dimension both at pretest and posttest. If there is no treatment effect and given the fan spread hypothesis the two correlations will be equal. This is because correlations remain invariant to linear transformations as in the fan spread hypothesis. A significant difference between the correlation for pretest and the correlation for posttest would suggest a significant treatment effect.

At present I see no clear advantages of this correlation comparison procedure over the estimated true score ANCOVA. Both procedures control

for confounding reflected in initial pretest differences as well as confounding that follows the fan spread hypothesis. However, the correlation comparison procedure requires that the design involve a pretest and a posttest both of identical form. Pretest and posttest of the same form are generally not available in the Follow Through data because as previously mentioned an achievement test appropriate for children starting kindergarten is often no longer appropriate for those children when they exit the program at third grade. If the pretest and posttest are not identical or at least parallel forms, there is no reason to expect their correlations with treatment group membership to be equal given no treatment effect. Thus a change in test form becomes confounded with treatment effect. Even where a design involves pretests and posttests of the same or parallel forms the correlation comparison procedure does not appear easily generalized to the complex designs typically called for in large scale quasi-experiments such as that illustrated in Figure 1. For example, how would one test the significance of an FT - NFT by approach interaction? Further, the results of the correlation comparison strategy are in the metric of correlation coefficients rather than in the metric of the dependent variable. In my opinion, the estimated true score ANCOVA strategy which affords confidence intervals around treatment mean differences that are in the dependent variable metric better facilitates decisions about the educational significance of results.

Using gain scores as dependent variables for analysis of variance (ANOVA) is another popular strategy for controlling confounding variables. However, the problem that tests must change over years within a cohort because of difficulty level, renders the use of gain scores inappropriate for the Follow Through evaluation. Even if the Follow Through evaluation did provide pretests and posttests of the same or parallel forms the use of gain scores involves another weakness. The hypothesis tested by gain scores is identical to the hypothesis tested by estimated true score ANCOVA except that the constant 1 replaces the slope of the regression line for the latent true parts of the dependent variable, Y, and the covariable, X. The comparison of two treatment groups becomes

$$\mu_{Y_{1.}} - \mu_{Y_{2.}} - (\mu_{X_{1.}} - \mu_{X_{2.}}) .$$

In the earlier example of the fan spread hypothesis the use of gain scores would have subtracted the initial two point difference from the final four point difference and concluded that there was a two point treatment difference rather than the correct conclustion of no treatment difference. Thus the use of gain scores corrects for initial differences but it does not control for the fan spread hypothesis.

Even for data that do not involve the type of confounding variables that happen during the course of a study, such as those which follow the fan spread hypothesis, the use of gain scores is not as good an analysis strategy as is estimated true score ANCOVA. For such data both procedures test the same hypothesis but estimated true scores ANCOVA does so with better precision. A detailed discussion of this point is not appropriate in the present paper. Briefly the better precision of estimated true scores ANCOVA results from the use of a substitute covariable followed by least squares, while gain scores use the observed variables with a non-least squares solution (Porter, paper in progress).

Another common strategy for controlling confounding variables is matching. As has already been indicated NFT schools were matched with FT schools in the Follow Through design. Although matching as a strategy to control confounding variables has received some deserved criticism (Campbell and Erlebacher, 1970), I think it can play an important role in the overall strategy for analysing quasi-experiments.

Again consider the two categories of confounding variables that are of concern in quasi-experiments, i.e., initial differences and differences that occur during the study other than those due to treatment. Matching is not as good a strategy for controlling initial differences as is estimated true scores ANCOVA for at least two reasons. First, effecting a good match of experimental units across treatment conditions on a pretest would require pretesting many more experimental units than would eventually be used. This would not be a problem if the pretests just happened to be administered to the population of interest at the desired time, but such

was not the case in the national evalution of Follow Through. The second reason is due to the source of internal invalidity that Campbell and Stanley (1963) label selection by regression interaction and which has been excellently described in a paper by Campbell and Erlebacher (1970). Consider two samples that have been selected so that they are matched on the basis of pretest scores, one from each of two populations having different means on the pretest. The sample selected from the population having the lower mean will tend to have a lower posttest mean than the sample selected from the other population. This is true despite a perfect match of the samples on the pretest unless the pretest is perfectly correlated with the posttest. In contrast, consider that a random sample was taken from each of the same two populations and the posttests were compared using estimated true scores on the pretest as the covariable in ANCOVA. The difference in pretest means for the samples would be an unbiased estimate of the initial population differences. As seen earlier in this paper using estimated true scores ANCOVA removes the estimate of initial differences from the posttest differences. Thus the analysis would tend not to show any spurious treatment effects.

In my opinion the real value of matching lies with its potential for controlling confounding variables which fall in the second category, particularly those which do not conform to the fan spread hypothesis and are therefore not controlled by estimated true scores ANCOVA. Although I believe pretests to be the best predictors of initial differences it does not necessarily follow that they are also the best predictors of differences that occur in the dependent variable dimension during program participation which are not a function of program participation. My reasoning is that initial differences are a function of all that has preceded the study in the life of the child, while differences that occur during the study other than due to program most likely are primarily a function of the child's environment at that time. For example, if socio-economic status of a family is related to ability of the home environment to effect changes in children that are tapped by the Follow Through test battery, then socio-economic status would be a good matching variable for

the Follow Through evaluation. If geographic location is related to the general quality of public school programs, then geographic location would be a good matching variable for the Follow Through evaluation.

The variables that appear to be the best predictors of differences due to other than treatments which occur while a study is being conducted may also be those for which an interaction with treatments is suspected. Matching on such variables guarantees that they will be crossed with treatment thus facilitating a test of the interaction. For example, if schools had been matched on ethnic composition across approaches in Figure 1, approach by ethnic composition interaction could have been tested rather than confounded with approaches as was earlier seen to be the case. Another advantage of matching to control for confounding variables is that the relationship between the matching variable and the dependent variable need not be linear for the match to be effective. To the contrary, one need not have any knowledge of the nature of the relationship, other than that it exists, in order that matching be effective.

When a design has employed matching primarily as a strategy for controlling confounding variables, it is sometimes forgotton in the analysis. Figure 1 illustrates that an attempt was made to match NFT schools with FT schools by representing schools, locations, and approaches as being crossed with FT - NFT. Any analysis must treat these sources of variation as crossed. The consequences of ignoring the matching by analyzing the data as though any of the crossed variables were nested, would be to violate the assumption of independence and in general to have tests of hypotheses about FT - NFT and approaches that are too conservative. For example, if locations were considered fixed and schools were analyzed as nested within FT - NFT, the schools nested within FT - NFT sum of squares would be equal to the sum of the sums of squares for schools and schools by FT - NFT interaction in the correct analysis. Since the FT - NFT sum of squares would be the same for both correct and incorrect analyses, the F test for and FT - NFT main effect using schools nested within FT - NFT would tend to have too large a denominator. This statement is partly off set by the sum of squares for schools nested within FT - NFT having more degrees of

freedom (the number of pairs of schools minus one) than the correct sum of squares error, i.e.,schools by FT - NFT interaction. However, the difference in degrees of freedom is not likely to be as large a factor as is the addition of the sum of squares for schools.

In conclusion, a combination of strategies employing both matching and estimated true score ANCOVA seems appropriate for attempting to control confounding variables in the national evaluation of Follow Through. The first step would be to choose a manageable number of variables that are not highly related to each other but which are believed to be predictive of differences that occur in the dependent variable dimensions during a cohorts participation in Follow Through and which are not a function of that participation. These variables might be selected by investigating the interrelationships of several potential variables and their relationships with gain scores for the variables in the Follow Through test battery. Where gain scores are not available perhaps some other index of change might be contrived or perhaps gain scores over only part of the total experience might be used. The variables most strongly related to change on the Follow Through test battery and with low interrelationships would be selected. The set of selected variables would then be used to match broad bands or levels of FT and NFT schools. In cases where a level contained more than enough NFT schools, a random sample would be taken. By taking a random sample from a pool of more than enough schools the problem of selection by regression interaction will be reduced. This is because regression only occurs when a sample is selected because it represents an extreme in a population distribution. After the matching process, estimated true scores would be calculated for the pretests and used as covariables in ANCOVA.

Unfortunately, matching after the data have been collected is limited to the extent that the distributions from which samples are to be matched are at least partially overlapping. For example, no match on ethnic composition of schools is possible for the two sponsors considered earlier, one of which had nearly all black children and the other of which had all white children. For some matching variables of interest the Follow Through design will afford a match across a few but not all approaches. For

example, if one wished to match on location there are a few places where more than one approach has been implemented, e.g.,New York City and Philadelphia. Separate analyses for each location could be done to compare approaches. Where two locations each have two approaches and one approach is in both locations, an indirect comparison of the other two approaches might be possible.

REFERENCES

Campbell, D. T. Temporal changes in treatment effect correlations: a quasi-experimental model for institutional records and longitudinal studies. In G. V. Glass (Ed.), Proceedings of the 1970 invitational conference on testing problems - The promise and perils of educational information systems.

Campbell, D.T. and Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), Compensatory Education: A national debate. Vol. III of The disadvantaged child, New York: Brunner/Mazel, 1970, 185-210.

Campbell, D.T. and Stanley, J.C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 171-246.

Cornfield, J. and Tukey, J. W. Average values of mean squares in factorials. Annals of Math. Stat., 1956, 27, 907-949.

Cox, D.R. The use of a concomitant variable in selecting an experimental design. Biometrika, 1957, 44, 150-158.

De Gracie, J. Analysis of covariance when the concomitant variable is measured with error. Ph.D. dissertation, Iowa State University, 1968.

Lord, F.M. Large - sample covariance analysis when the control variable is fallible. J. Amer. Statist. Assn., 1960, 55, 307-321.

Porter, A.C. The effects of using fallible variables in the analysis of covariance. Ph.D. dissertation, University of Wisconsin, 1967.

Porter, A.C. Comments on some current strategies to evaluate the effectiveness of compensatory education programs. A paper presented at the Annual Meeting of the American Psychological Association, 1969.

Porter, A.C. How errors of measurement affect ANOVA, regression analyses, ANCOVA and factor analyses. A paper presented at the American Educational Research Association meetings, 1971.

Porter, A.C. Analysis strategies for some common evaluation paradigms. Paper in progress.

Smith, M.S. and Bissell, Joan S. Report analysis: The impact of Head Start. Harvard Educational Review, 1970, 40, No. 1, 51-104.

U.S. Office of Education Request for proposal to analyze the national evaluation of Follow Through data. RFP 72-37, April, 1972.

Wright, D.J. Groups and experimental units in educational research. A paper presented at the American Educational Research Association meetings, 1969.